

Origin of the computational hardness for learning with binary synapses

Haiping Huang and Yoshiyuki Kabashima

*Department of Computational Intelligence and Systems Science,
Tokyo Institute of Technology, Yokohama 226-8502, Japan*

(Dated: August 11, 2014)

Supervised learning in a binary perceptron is able to classify an extensive number of random patterns by a proper assignment of binary synaptic weights. However, to find such assignments in practice, is quite a nontrivial task. The relation between the weight space structure and the algorithmic hardness has not yet been fully understood. To this end, we analytically derive the Franz-Parisi potential for the binary perceptron problem, by starting from an equilibrium solution of weights and exploring the weight space structure around it. Our result reveals the geometrical organization of the weight space—the weight space is composed of isolated solutions, rather than clusters of exponentially many close-by solutions. The point-like clusters far apart from each other in the weight space explain the previously observed glassy behavior of stochastic local search heuristics.

PACS numbers: 89.75.Fb, 87.19.L-, 75.10.Nr

I. INTRODUCTION

To provide an analytic explanation for general phenomena using simple theoretical concept is the most interesting part of physics. Statistical physics methods in spin glass theory provide new tools and ideas to study many hard constraint satisfaction problems [1], especially the relation between detailed organization of solutions in the solution space and the algorithmic hardness [2].

A prototypical example is the binary perceptron problem, where N input neurons (units) are connected to a single output unit by synapses of binary value (± 1) synaptic weights. These weights have to be inferred from a set of examples (input patterns) with desired classification labels (supervised learning). An assignment of these weights is referred to as a solution if the perceptron manages to classify all the input patterns by this assignment. The ratio between the number of patterns and the number of synapses is called the constraint density. Each example acts as a constraint on the solution space, since increasing the number of examples causes the shrinkage of the space. The critical constraint density was reported to be about 0.833 [3], below which the solution space is typically nonempty.

The binary perceptron serves as an elementary building block of complex neural networks and is also one of the basic structures for learning and memory [4]. Memory in neuronal systems is stored in the synaptic weights, and a binary synaptic weight is robust against noise and also suitable for simple hardware implementation in applications. The binary perceptron has thus a wide variety of applications ranging from rule inference or structure mining in machine learning [4] to error correcting codes or data compression in information theory [5], and even high-dimensional data analysis in biology [6]. However, a learning task in the binary perceptron is known to be an NP(nondeterministic polynomial time)-complete problem in the worst case [7]. Many efforts have been devoted to design low-complexity algorithms to find a solution for a typical case of this difficult problem [8–15]. However, for many local search heuristics, the search process slows down as the constraint density grows, and the learning threshold decreases as the number of synapses increases [9, 13, 14]. This typical glassy behavior of stochastic local search algorithms remains to be explained and was conjectured to be related to the geometrical organization of the solution space [14, 16–18]. The statistical properties of this problem were intensively studied by the statistical physics community in the past decades [3, 4, 17, 19]. However, an analytic computation of a conclusive picture of the solution space structure is still lacking so far, although this is an important topic both in computer science (machine learning or computational neuroscience) and in statistical physics.

A recent study [18] carried out an entropy landscape analysis by focusing on the solution-pairs separated by certain Hamming distance (the number of elements in different states in two solutions), which motivated us to propose a suitable and solid framework to provide a comprehensive description of the solution space. The basic idea is to select an equilibrium solution sampled from the Boltzmann measure, and then explore the solution space around this selected equilibrium solution by analyzing the entropy landscape in the vicinity of the reference equilibrium solution. This framework was originally introduced as the name of Franz-Parisi potential to study the metastable state structure for discontinuous mean-field spin glasses (e.g., p -spin spherical spin glass) [20–22], where the potential has the physical meaning of the free energy cost to keep a system at a temperature with a fixed overlap from an equilibrium configuration at a different temperature. In this work, the Franz-Parisi potential is interpreted in terms of the entropy function to describe the solution space, and we show that a quenched computation (average over the choice of the reference equilibrium solution) of the potential in the zero temperature limit is possible and provides important physical insights towards understanding the geometrical organization of the solution (weight) space.

Our computation demonstrates that the weight space of the binary perceptron problem is indeed made of isolated solutions for any finite constraint density, with the minimal Hamming distance separating two solutions growing with the constraint density. This study reveals the origin of the computational hardness in the binary perceptron problem, explaining the known fact that when the number of synapses becomes sufficiently large, an exponential scaling in computational time is required to maintain a fixed finite constraint density for a learning task [9, 14, 16].

In Sec. II, we define in detail the binary perceptron problem. In Sec. III, we introduce the Franz-Parisi potential framework and derive the explicit form of the potential under the replica symmetric approximation. Results are presented and discussed in Sec. IV. Concluding remarks and future perspectives are given in Sec. V.

II. THE BINARY PERCEPTRON PROBLEM

The binary perceptron is a single-layered feed-forward neural network, i.e., N input neurons are connected to a single output neuron by N synapses of weight $J_i = \pm 1$ ($i = 1, 2, \dots, N$). The perceptron tries to learn $P = \alpha N$ associations $\{\xi^\mu, \sigma_0^\mu\}$ ($\mu = 1, 2, \dots, P$), where $\xi^\mu \equiv (\xi_1^\mu, \xi_2^\mu, \dots, \xi_N^\mu)$ is an input pattern with $\xi_i^\mu = \pm 1$, and $\sigma_0^\mu = \pm 1$ is the desired classification of the input pattern μ . For a random classification task, both $\{\xi_i^\mu\}$ and the desired output $\{\sigma_0^\mu\}$ are generated randomly independently with ξ_i^μ and σ_0^μ being ± 1 with probability $1/2$. Given the input pattern ξ^μ , the actual output σ^μ of the perceptron is $\sigma^\mu = \text{sgn} \left(\sum_{i=1}^N J_i \xi_i^\mu \right)$. If $\sigma^\mu = \sigma_0^\mu$, we say that the synaptic weight vector \mathbf{J} has learned the μ -th pattern. Each input pattern imposes a constraint on all synaptic weights, therefore α denotes the constraint density. The solution space of the binary perceptron is composed of all the weight configurations $\{J_i\}$ that satisfy $\sigma_0^\mu \sum_i J_i \xi_i^\mu > 0$ for $\mu = 1, 2, \dots, P$. The energy cost is thus defined as the number of patterns mapped incorrectly [4, 18], i.e.,

$$E(\mathbf{J}) = \sum_{\mu} \Theta \left(-\frac{\sigma_0^\mu}{\sqrt{N}} \sum_{i=1}^N J_i \xi_i^\mu \right), \quad (1)$$

where $\Theta(x)$ is a step function with the convention that $\Theta(x) = 0$ if $x \leq 0$ and $\Theta(x) = 1$ otherwise. The prefactor $N^{-1/2}$ is introduced to ensure that the argument of the step function remains at the order of unity, for the sake of the following statistical mechanical analysis in the thermodynamic limit. Without loss of generality, we assume $\sigma_0^\mu = +1$ for any input pattern in the remaining part of this paper, since one can perform a gauge transformation $\xi_i^\mu \rightarrow \xi_i^\mu \sigma_0^\mu$ to each input pattern without affecting the result.

From a theoretical perspective, the perceptron is typically able to learn an extensive number of random input patterns with the storage capacity $\alpha_s \simeq 0.833$ [3]. However, to find such a solution configuration \mathbf{J} in practice, is quite a nontrivial task. Here, to reveal the origin of this computational hardness, we apply the replica method from the theory of disordered systems [1] to derive an analytic expression of the Franz-Parisi potential, which characterizes the entropy landscape of the problem.

III. ANALYTIC COMPUTATION OF THE FRANZ-PARISI POTENTIAL

The binary perceptron problem is a densely-connected graphical model [18] in that a proper assignment of all synaptic weights is needed to satisfy each constraint (learn each pattern). Its equilibrium property can thus be described by mean-field computation in terms of the Franz-Parisi potential. The basic idea is to first select an equilibrium configuration \mathbf{J} at a temperature T' , then constrain its overlap with another equilibrium configuration \mathbf{w} at a different temperature T , which yields a constrained free energy [20]:

$$F(T, T', x) = \left\langle \frac{1}{Z(T')} \sum_{\mathbf{J}} e^{-\beta' E(\mathbf{J})} \ln \sum_{\mathbf{w}} e^{-\beta E(\mathbf{w}) + x \mathbf{J} \cdot \mathbf{w}} \right\rangle, \quad (2)$$

after taking the quenched disorder average (over the pattern distribution ξ , denoted by the angular bracket) and the average over the distribution of \mathbf{J} , which is $e^{-\beta' E(\mathbf{J})}/Z(T')$. $Z(T')$ is the partition function for the original measure and $\beta(\beta')$ is the inverse temperature. The constrained free energy $\ln \sum_{\mathbf{w}} e^{-\beta E(\mathbf{w}) + x \mathbf{J} \cdot \mathbf{w}}$ is a self-averaging quantity with respect to both the quenched disorder and the probability distribution of the reference configuration \mathbf{J} [21]. Its value does not depend on the particular realization and coincides with the typical value, which can be calculated via the replica method.

In our current setting, we are interested in the ground states of the problem, thus we set $\beta = \beta' \rightarrow \infty$, arriving at the following formula:

$$F(x) = \lim_{\substack{n \rightarrow 0 \\ m \rightarrow 0}} \frac{\partial}{\partial m} \left\langle \sum_{\{\mathbf{J}^a, \mathbf{w}^\gamma\}} \prod_{\mu} \left[\prod_{a, \gamma} \Theta(u_a^\mu) \Theta(v_\gamma^\mu) \right] e^{x \sum_{\gamma, i} J_i^1 w_i^\gamma} \right\rangle, \quad (3)$$

where $u_a^\mu \equiv \sum_i J_i^a \xi_i^\mu / \sqrt{N}$ and $v_\gamma^\mu \equiv \sum_i w_i^\gamma \xi_i^\mu / \sqrt{N}$. In Eq. (A3), we have n replicas \mathbf{J}^a ($a = 1, \dots, n$) and m replicas \mathbf{w}^γ ($\gamma = 1, \dots, m$), with the coupling field (x) term being an interaction of all the constrained replicas \mathbf{w}^γ with one privileged replica \mathbf{J}^1 . The replica method to compute the typical value of the constrained free energy is based on two mathematical identities: $\ln Z = \lim_{m \rightarrow 0} \frac{\partial Z^m}{\partial m}$ and $Z^{-1} = \lim_{n \rightarrow 0} Z^{n-1}$. To evaluate the average in Eq. (A3), we need to define the overlap matrixes $Q_{ab} \equiv \mathbf{J}^a \cdot \mathbf{J}^b / N$, $P_{a\gamma} \equiv \mathbf{J}^a \cdot \mathbf{w}^\gamma / N$ and $R_{\gamma\eta} \equiv \mathbf{w}^\gamma \cdot \mathbf{w}^\eta / N$, which characterize the following disorder averages $\langle u_a^\mu u_b^\mu \rangle = Q_{ab}$, $\langle u_a^\mu v_\gamma^\mu \rangle = P_{a\gamma}$ and $\langle v_\gamma^\mu v_\eta^\mu \rangle = R_{\gamma\eta}$. Under the replica symmetric (RS) ansatz, we have $Q_{ab} = q(1 - \delta_{ab}) + \delta_{ab}$, $P_{a\gamma} = p\delta_{a1} + p'(1 - \delta_{a1})$ and $R_{\gamma\eta} = r(1 - \delta_{\gamma\eta}) + \delta_{\gamma\eta}$, where $\delta_{ab} = 1$ if $a = b$ and 0 otherwise.

After some algebraic manipulations, we finally get the constrained free energy density $f(x)$ as:

$$f(x) = \lim_{N \rightarrow \infty} F(x)/N = \frac{\hat{r}}{2}(r-1) - p\hat{p} + p'\hat{p}' + xp + \alpha \int D\omega \int Dt H^{-1}(\tilde{t}) \int_{\tilde{t}}^\infty Dy \ln H(h(\omega, t, y)) \\ + \int D\mathbf{z} (2 \cosh \hat{a})^{-1} \left[e^{\hat{a}} \ln 2 \cosh(\hat{a}' + \hat{p} - \hat{p}') + e^{-\hat{a}} \ln 2 \cosh(\hat{a}' - \hat{p} + \hat{p}') \right], \quad (4)$$

where $\int D\mathbf{z} \equiv \int Dz_1 Dz_2 Dz_3$, $\tilde{t} \equiv -\sqrt{\frac{q}{1-q}}t$, and $H(x) \equiv \int_x^\infty Dz$ with the Gaussian measure $Dz \equiv G(z)dz$ in which $G(z) \equiv \exp(-z^2/2)/\sqrt{2\pi}$. $h(\omega, t, y) \equiv -((p-p')y/\sqrt{1-q} + \sqrt{v_\omega}\omega + p't/\sqrt{q})/\sqrt{1-r}$ where $v_\omega \equiv r - p'^2/q - (p-p')^2/(1-q)$. $\hat{a} \equiv \sqrt{\hat{q} - \hat{p}'}z_1 + \sqrt{\hat{p}'}z_3$ and $\hat{a}' \equiv \sqrt{\hat{r} - \hat{p}'}z_2 + \sqrt{\hat{p}'}z_3$. The associated self-consistent (saddle-point) equations for the order parameters $\{q, \hat{q}, r, \hat{r}, p, \hat{p}, p', \hat{p}'\}$ are derived in the Appendix A.

The Franz-Parisi potential $\mathcal{V}(p)$ is obtained through a Legendre transform of $f(x)$, i.e., $\mathcal{V}(p) = f(x) - xp$ and $\frac{df(x)}{dx} = p$. $\mathcal{V}(p)$ has the meaning of the entropy characterizing the growth rate of the number of solutions ($e^{N\mathcal{V}(p)}$) lying apart at a normalized distance $(1-p)/2$ (Hamming distance divided by N) from the fixed equilibrium solution. Detailed information about the solution space structure can be extracted from the behavior of this potential at different values of p , especially those values close to one. Since the potential curve may lose its concavity, one has to solve numerically the saddle-point equations (see Appendix B) by fixing p and searching for compatible coupling field x (by using the secant method [23]).

IV. SOLUTION SPACE CONSISTS OF ISOLATED SOLUTIONS

The Franz-Parisi potential versus the predefined normalized Hamming distance ($d = (1-p)/2$) is shown in Fig. 1 (a). At the maximum corresponding to $x = 0$ ($x = -\frac{d\mathcal{V}}{dp} = \frac{1}{2}\frac{d\mathcal{V}}{dd}$), $\mathcal{V}(p)$ gives back the entropy of the original system. As the distance gets close to zero, one finds that there exists a value of distance at which the entropy curve loses its concavity and turns to a convex part (see the inset of Fig. 1 (a) and note that the sign of the slope changes at the maximum point). This behavior leads to an important result that there exists a minimal distance of $\mathcal{O}(N)$ below which no solutions are separated from the reference equilibrium solution. Note that the reference solution is distributed according to the Boltzmann measure (a uniform measure over all solutions). The minimal distance grows with the constraint density, as shown in Fig. 1 (b). This can be understood by the following argument. Due to the hard nature of the pattern constraint in the binary perceptron problem—all synapses are involved in classifying each input pattern, flipping one synaptic weight should force the rearrangement of many weight values to memorize the learned patterns. Similar phenomena were also observed in Gallager's type error correcting code [24] and locked constraint satisfaction problem [25].

For small α , it is not easy to show the convex part numerically. However, one can prove that when $p \rightarrow 1$, the Franz-Parisi potential vanishes as expected for all α (see Appendix C). In addition, at $p \rightarrow 1$ ($\epsilon \equiv 1-p \rightarrow 0$), we have $\frac{d\mathcal{V}(p)}{dp} = \alpha C_p \epsilon^{-1/2} + (\ln \epsilon)/2 + C$ (see Appendix B) where C is a finite constant and C_p is a positive constant. The first term dominates the divergent behavior in the limit $\epsilon \rightarrow 0$. This means that, for any finite $\alpha > 0$, the entropy curve in Fig. 1 (a) has a negative infinite slope ($\frac{d\mathcal{V}}{dd} = -2\frac{d\mathcal{V}}{dp}$) at $p = 1$, supporting the existence of the convex part and the minimal distance. As expected from the tendency shown in Fig. 1 (b), the value of the minimal distance

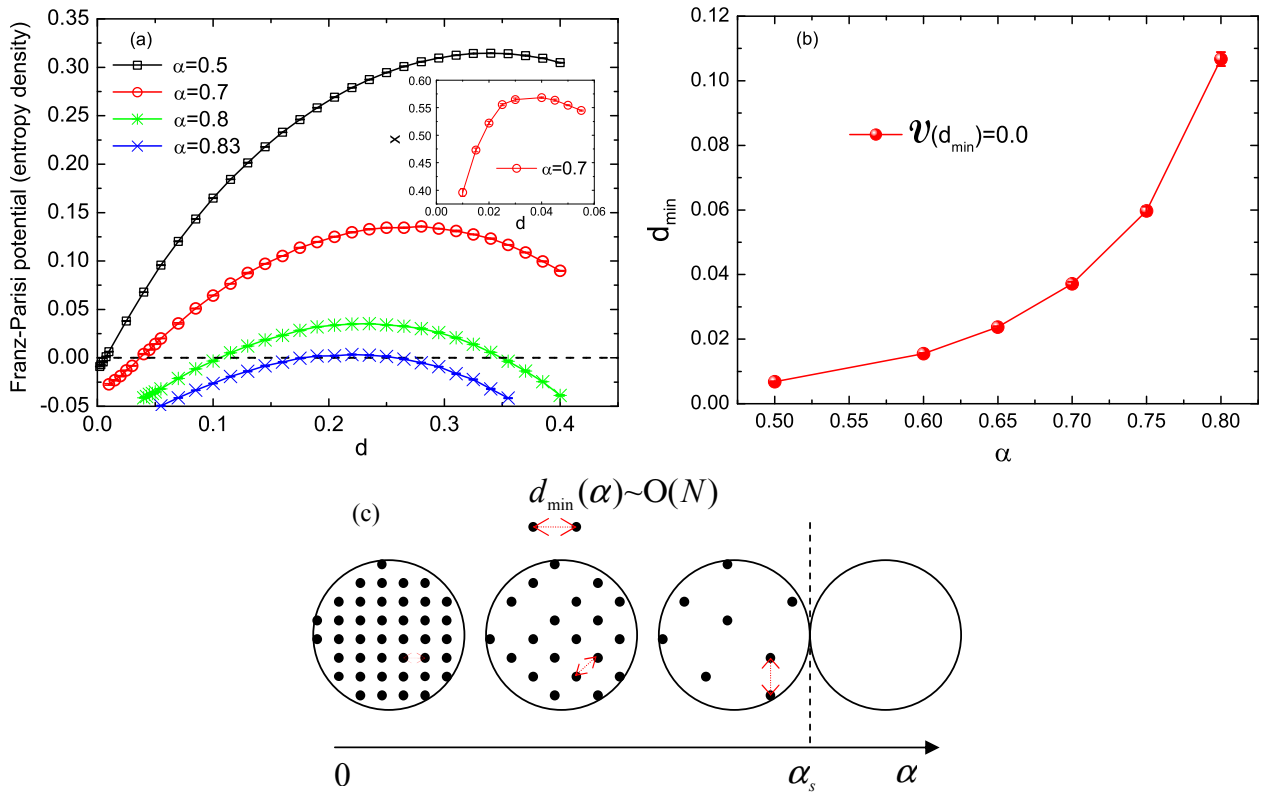


FIG. 1: (Color online) Entropy landscape of solutions in the binary perceptron problem. Iterations of the saddle-point equations are always converged to produce the data points. The error bars give statistical errors and are smaller than or equal to the symbol size. (a) Franz-Parisi potential as a function of the normalized Hamming distance. The behavior of the coupling field with the distance is shown in the inset for $\alpha = 0.7$, for which an observed maximum implies the change of the concavity of the entropy curve (this also holds for other finite values of α). (b) Minimal distance versus the constraint density. Within the minimal distance, there are no solutions satisfying the distance constraint from the reference equilibrium solution. (c) Schematic illustration of the weight space based on results of (a) and (b). The points indicate the equilibrium solutions of weights. $\alpha_s \simeq 0.833$ is the storage capacity after which the solution space is typically empty. d_{\min} is the actual Hamming distance without normalization.

becomes very small for the less constrained case (small constraint density). This explains why a simple local search algorithm can find a solution when either N or α is small [8–10, 13–15]. As α increases, the minimal distance grows rapidly, as a consequence, any algorithms working by local move (each time a few weights are flipped) should find increasing difficulty to identify a solution (especially at a very large N), which holds even for reinforced message passing algorithms [11]. In other words, an extensive energy or entropic barrier should be overcome. The energy landscape is always valleys dominated (valleys are metastable states with positive energy cost). These metastable states are much more numerous than the frozen ground states [26]. Local algorithms will get trapped by these metastable states with high probability.

We thus conclude that, at variance with random K -SAT or Q -coloring problems [2], the solution space of the binary perceptron problem is simple in the sense that it is made of isolated solutions instead of well separated clusters of exponentially many close-by solutions. This picture is consistent with evidences reported in previous studies [17, 18, 27]. Moreover, non-convergence of the iteration of the saddle-point equations was not observed, which may be related to the simple structure of the solution space. In fact, below the storage capacity, the replica symmetric solution is stable without any need to introduce replica symmetry breaking scenario for this problem [3, 19]. Our quenched computation of the Franz-Parisi potential reveals that, synaptic weights to realize the random classification task are organized into point-like clusters (zero internal entropy) far apart from each other (see Fig. 1 (c)), with the result that in the thermodynamic limit, an exponential computation time is required to reach a finite fixed α [9, 16].

V. CONCLUSION

We give an analytic expression of the Franz-Parisi potential for the binary perceptron problem. This potential describes the entropy landscape of solutions in the vicinity of a reference equilibrium solution, and its shape is independent of the choice of the reference point. Solving the saddle-point equations, we find that the concavity of the curve changes at some distance, leading to a minimal distance below which there does not exist solutions satisfying the distance constraint. Furthermore, this minimal distance increases with the constraint density, implying that the problem is extremely hard because the solution space is composed of isolated solutions (point-like clusters) with the property that to go from one solution to another solution, one should flip an extensive number (proportional to N) of synaptic weights.

Our analysis establishes a refined picture of the organization structure of the solution space for the binary perceptron problem, which is helpful for understanding the glassy behavior of local search heuristics [9, 13, 14], which may have some connections with recent studies of constrained glasses [28], and furthermore, is expected to shed light on design of efficient algorithms for large-scale neuromorphic devices. The analytic analysis presented in this paper also offers a basis for possible rigorous mathematical (probabilistic) analysis of the entropy landscape [29], and has potentially applications for studying the solution space structure of other hard problems in information processing, e.g., spike time-based neural classifiers [30–32].

Acknowledgments

We thank Lenka Zdeborová for helpful discussions and Haijun Zhou for helpful comments on the manuscript. This work was partially supported by the JSPS Fellowship for Foreign Researchers (Grant No. 24·02049) (H.H.) and JSPS/MEXT KAKENHI Grant No. 25120013 (Y.K.). Support from the JSPS Core-to-Core Program “Non-equilibrium dynamics of soft matter and information” is also acknowledged.

Appendix A: Derivation of constrained free energy

In the current context, for a reference equilibrium configuration \mathbf{J} at temperature T' , one is interested in the free energy of a perturbed system (with the constraint that the configuration \mathbf{w} at temperature T should satisfy a prefixed overlap with \mathbf{J}), leading to the constrained free energy [20]:

$$F(T, T', x) = \left\langle \frac{1}{Z(T')} \sum_{\mathbf{J}} e^{-\beta' E(\mathbf{J})} \ln \sum_{\mathbf{w}} e^{-\beta E(\mathbf{w}) + x \mathbf{J} \cdot \mathbf{w}} \right\rangle_{\xi}, \quad (\text{A1})$$

where $Z(T') = \sum_{\mathbf{J}} e^{-\beta' E(\mathbf{J})}$ and x is the coupling field to control the overlap (or distance) between two configurations, i.e., $p \equiv \mathbf{J} \cdot \mathbf{w}/N$. We are interested in the ground state, then we set both inverse temperatures equal and make them tend to infinity. Substituting the definition of energy cost of the problem, and using $e^{-\beta \Theta(-u)} = \Theta(u)$ in the zero temperature limit, we have

$$F(x) = \left\langle \frac{1}{Z(T')} \sum_{\mathbf{J}} \Theta \left(\frac{1}{\sqrt{N}} \sum_{i=1}^N J_i \xi_i^\mu \right) \ln \sum_{\mathbf{w}} \Theta \left(\frac{1}{\sqrt{N}} \sum_{i=1}^N w_i \xi_i^\mu \right) e^{x \mathbf{J} \cdot \mathbf{w}} \right\rangle_{\xi}. \quad (\text{A2})$$

To evaluate the typical value of $F(x)$, we resort to the replica method [4], by using two mathematical identities: $\ln Z = \lim_{m \rightarrow 0} \frac{\partial Z^m}{\partial m}$ and $Z^{-1} = \lim_{n \rightarrow 0} Z^{n-1}$. Introducing n unconstrained replicas $\mathbf{J}^a (a = 1, \dots, n)$ and m constrained replicas $\mathbf{w}^\gamma (\gamma = 1, \dots, m)$, we rewrite $F(x)$ as:

$$F(x) = \lim_{\substack{n \rightarrow 0 \\ m \rightarrow 0}} \frac{\partial}{\partial m} \left\langle \sum_{\{\mathbf{J}^a, \mathbf{w}^\gamma\}} \prod_{\mu} \left[\prod_{a, \gamma} \Theta(u_a^\mu) \Theta(v_\gamma^\mu) \right] e^{x \sum_{\gamma, i} J_i^1 w_i^\gamma} \right\rangle_{\xi}, \quad (\text{A3})$$

where $u_a^\mu \equiv \sum_i J_i^a \xi_i^\mu / \sqrt{N}$ and $v_\gamma^\mu \equiv \sum_i w_i^\gamma \xi_i^\mu / \sqrt{N}$. To proceed, we define the following overlap matrixes: $Q_{ab} \equiv \mathbf{J}^a \cdot \mathbf{J}^b / N$, $P_{a\gamma} \equiv \mathbf{J}^a \cdot \mathbf{w}^\gamma / N$ and $R_{\gamma\eta} \equiv \mathbf{w}^\gamma \cdot \mathbf{w}^\eta / N$, which characterize the following disorder averages $\langle u_a^\mu u_b^\mu \rangle = Q_{ab}$, $\langle u_a^\mu v_\gamma^\mu \rangle = P_{a\gamma}$ and $\langle v_\gamma^\mu v_\eta^\mu \rangle = R_{\gamma\eta}$. By inserting delta functions for these definitions and using their integral

representations, we obtain the disorder average \mathcal{S} in Eq. (A3) as:

$$\begin{aligned} \mathcal{S} = & \prod_{a < b} \prod_{\gamma < \eta} \prod_{a, \gamma} \int \frac{dQ_{ab} d\hat{Q}_{ab}}{2\pi} \int \frac{dR_{\gamma\eta} d\hat{R}_{\gamma\eta}}{2\pi} \int \frac{dP_{a\gamma} d\hat{P}_{a\gamma}}{2\pi} e^{-i(\sum_{a < b} Q_{ab} \hat{Q}_{ab} + \sum_{\gamma < \eta} R_{\gamma\eta} \hat{R}_{\gamma\eta} + \sum_{a, \gamma} P_{a\gamma} \hat{P}_{a\gamma})} \\ & \times \sum_{\{\mathbf{J}^a, \mathbf{w}^\gamma\}} e^{\frac{i}{N}(\sum_{a < b} \hat{Q}_{ab} \sum_i J_i^a J_i^b + \sum_{\gamma < \eta} \hat{R}_{\gamma\eta} \sum_i w_i^\gamma w_i^\eta + \sum_{a, \gamma} \hat{P}_{a\gamma} \sum_i J_i^a w_i^\gamma)} \\ & \times \left\langle \prod_{\mu} \left[\prod_{a, \gamma} \Theta(u_a^\mu) \Theta(v_\gamma^\mu) \right] \right\rangle_{\xi} e^{x \sum_{i, \gamma} J_i^a w_i^\gamma}. \end{aligned} \quad (\text{A4})$$

Now we re-scale the variable $i\hat{Q}_{ab}/N \rightarrow \hat{Q}_{ab}$ (this also applies for other conjugated variables). We apply the replica symmetric approximation [4], which assumes the permutation symmetry of the overlap matrix. To be more precise, $Q_{ab} = q(1 - \delta_{ab}) + \delta_{ab}$, $P_{a\gamma} = p\delta_{a1} + p'(1 - \delta_{a1})$ and $R_{\gamma\eta} = r(1 - \delta_{\gamma\eta}) + \delta_{\gamma\eta}$, where $\delta_{ab} = 1$ if $a = b$ and 0 otherwise. We first simplify $\sum_{a, \gamma} \hat{P}_{a\gamma} J^a w^\gamma$ as:

$$\begin{aligned} \sum_{a, \gamma} \hat{P}_{a\gamma} J^a w^\gamma &= \hat{p}' \sum_{a, \gamma} J^a w^\gamma + (\hat{p} - \hat{p}') \sum_{\gamma} J^1 w^\gamma \\ &= \frac{\hat{p}'}{2} \left[\left(\sum_a J^a + \sum_{\gamma} w^\gamma \right)^2 - \left(\sum_a J^a \right)^2 - \left(\sum_{\gamma} w^\gamma \right)^2 \right] + (\hat{p} - \hat{p}') \sum_{\gamma} J^1 w^\gamma, \end{aligned} \quad (\text{A5})$$

where the site index i is dropped off since each i shares the same formula. Then we compute the disorder average as:

$$\left\langle \prod_{\mu} \left[\prod_{a, \gamma} \Theta(u_a^\mu) \Theta(v_\gamma^\mu) \right] \right\rangle_{\xi} = \left[\int D\omega \int Dt \int_{\tilde{t}}^{\infty} Dy H^m(h(\omega, t, y)) H^{n-1}(\tilde{t}) \right]^{\alpha N}, \quad (\text{A6})$$

where $\tilde{t} \equiv -\sqrt{\frac{q}{1-q}}t$, and $H(x) \equiv \int_x^{\infty} Dz$ with the Gaussian measure $Dz \equiv G(z)dz$ in which $G(z) = \exp(-z^2/2)/\sqrt{2\pi}$. $h(\omega, t, y) \equiv -((p-p')y/\sqrt{1-q} + \sqrt{v_\omega}\omega + p't/\sqrt{q})/\sqrt{1-r}$ where $v_\omega \equiv r - p'^2/q - (p-p')^2/(1-q)$. In deriving Eq. (A6), we have parameterized $u_a = \sqrt{1-q}y_a + \sqrt{qt}$ and $v_\gamma = \sqrt{1-r}y'_\gamma + (p-p')y_1/\sqrt{1-q} + \sqrt{v_\omega}\omega + p't/\sqrt{q}$, by using independent standard Gaussian random variables $\{y_a, t, y'_\gamma, \omega\}$ of zero mean and unit variance. The parameterization retains the covariance structure of $\{u_a, v_\gamma\}$. The pattern index (μ) is also dropped off for the same reason. After a few algebraic manipulations, we obtain

$$\begin{aligned} \mathcal{S} = & \exp \left[-\frac{N(n-1)n}{2} q\hat{q} - \frac{N(m-1)m}{2} r\hat{r} - mNp\hat{p} - N(n-1)mp'\hat{p}' + Nxpm - \frac{Nn}{2}\hat{q} - \frac{Nm}{2}\hat{r} \right] \\ & \times \exp \left[N \ln \int Dz_1 \int Dz_2 \int Dz_3 \mathcal{A}(\hat{q}, \hat{r}, \hat{p}, \hat{p}', m, n) \right] \\ & \times \exp \left[\alpha N \ln \int D\omega \int Dt \int_{\tilde{t}}^{\infty} Dy H^m(h(\omega, t, y)) H^{n-1}(\tilde{t}) \right], \end{aligned} \quad (\text{A7})$$

after approximating the integral in Eq. (A4) by its dominant part (a saddle point analysis in the large N limit). To derive Eq. (A7), the Hubbard-Stratonovich transformation was used. In Eq. (A7), $\mathcal{A}(\hat{q}, \hat{r}, \hat{p}, \hat{p}', m, n) \equiv (2 \cosh \hat{a})^{n-1} \left[e^{\hat{a}} (2 \cosh(\hat{a}' + \hat{p} - \hat{p}'))^m + e^{-\hat{a}} (2 \cosh(\hat{a}' - \hat{p} + \hat{p}'))^m \right]$, in which $\hat{a} \equiv \sqrt{\hat{q} - \hat{p}'z_1} + \sqrt{\hat{p}'z_3}$ and $\hat{a}' \equiv \sqrt{\hat{r} - \hat{p}'z_2} + \sqrt{\hat{p}'z_3}$.

The saddle point analysis (also called Laplace method) implies that \mathcal{S} should take its maximal value so that $\mathcal{L} \equiv \ln \mathcal{S}$ should be extremized with respect to the order parameters $\{q, \hat{q}, r, \hat{r}, p, \hat{p}, p', \hat{p}'\}$. Keeping up to the first order in n , the extremization with respect to q and \hat{q} gives the self-consistent equations for q and \hat{q} (see Eqs. (A9a) and (A9b)). As expected, their values do not rely on other order parameters characterizing the property of the constrained replicas. These two equations describe the \mathbf{J} system at equilibrium, and it should not be affected by the \mathbf{w} system which follows a perturbed distribution depending on the reference solution \mathbf{J} . Finally, one can readily get the constrained

free energy density following the definition given in Eq. (A3):

$$f(x) = \lim_{N \rightarrow \infty} F(x)/N = \frac{\hat{r}}{2}(r-1) - p\hat{p} + p'\hat{p}' + xp + \alpha \int D\omega \int Dt H^{-1}(\tilde{t}) \int_{\tilde{t}}^{\infty} Dy \ln H(h(\omega, t, y)) \\ + \int D\mathbf{z} (2 \cosh \hat{a})^{-1} \left[e^{\hat{a}} \ln 2 \cosh(\hat{a}' + \hat{p} - \hat{p}') + e^{-\hat{a}} \ln 2 \cosh(\hat{a}' - \hat{p} + \hat{p}') \right], \quad (\text{A8})$$

together with the associated saddle-point equations:

$$q = \int Dz \tanh^2(\sqrt{\hat{q}}z), \quad (\text{A9a})$$

$$\hat{q} = \frac{\alpha}{1-q} \int Dt \mathcal{R}^2(\tilde{t}), \quad (\text{A9b})$$

$$p = \int D\mathbf{z} (2 \cosh \hat{a})^{-1} \left[e^{\hat{a}} \tanh(\hat{a}' + \hat{p} - \hat{p}') - e^{-\hat{a}} \tanh(\hat{a}' - \hat{p} + \hat{p}') \right], \quad (\text{A9c})$$

$$\hat{p} = x + \frac{\alpha}{\sqrt{(1-q)(1-r)}} \int D\omega \int Dt \mathcal{R}(\tilde{t}) \mathcal{R}(h(\omega, t, y = \tilde{t})), \quad (\text{A9d})$$

$$r = \int D\mathbf{z} (2 \cosh \hat{a})^{-1} \left[e^{\hat{a}} \tanh^2(\hat{a}' + \hat{p} - \hat{p}') + e^{-\hat{a}} \tanh^2(\hat{a}' - \hat{p} + \hat{p}') \right], \quad (\text{A9e})$$

$$\hat{r} = \frac{\alpha}{1-r} \int D\omega \int Dt H^{-1}(\tilde{t}) \int_{\tilde{t}}^{\infty} Dy \mathcal{R}^2(h(\omega, t, y)), \quad (\text{A9f})$$

$$p' = \int D\mathbf{z} (2 \cosh \hat{a})^{-1} \left[e^{\hat{a}} \tanh \hat{a} \tanh(\hat{a}' + \hat{p} - \hat{p}') + e^{-\hat{a}} \tanh \hat{a} \tanh(\hat{a}' - \hat{p} + \hat{p}') \right], \quad (\text{A9g})$$

$$\hat{p}' = \frac{\alpha}{\sqrt{(1-q)(1-r)}} \int D\omega \int Dt H^{-1}(\tilde{t}) \mathcal{R}(\tilde{t}) \int_{\tilde{t}}^{\infty} Dy \mathcal{R}(h(\omega, t, y)), \quad (\text{A9h})$$

where $\int D\mathbf{z} \equiv \int Dz_1 Dz_2 Dz_3$, and $\mathcal{R}(x) \equiv G(x)/H(x)$. In deriving these equations, we have used a useful property of the Gaussian measure $\int Dz z \mathcal{F}(z) = \int Dz \mathcal{F}'(z)$ where $\mathcal{F}'(z)$ is the derivative of the function $\mathcal{F}(z)$ with respect to z .

To solve these saddle-point equations, for example, Eq. (A9f), one efficient way is to generate a random number y according to the conditional distribution $\Pr(y|t) = \frac{G(y)\Theta(\sqrt{1-q}y + \sqrt{q}t)}{H(-\sqrt{\frac{q}{1-q}}t)}$ each time when using Monte-Carlo method to perform the integral. In some cases, one may reexpress \hat{a} and \hat{a}' to retain their covariances $\langle \hat{a}\hat{a}' \rangle = \hat{p}'$ (their means are both zero, and variances $\langle \hat{a}^2 \rangle = \hat{q}, \langle \hat{a}'^2 \rangle = \hat{r}$) according to their definition, this is because, $\hat{q} - \hat{p}'$ or $\hat{r} - \hat{p}'$ may get negative.

Appendix B: Derivation of $\frac{d\mathcal{V}(p)}{dp}|_{p \rightarrow 1}$

The Franz-Parisi potential $\mathcal{V}(p)$ is obtained through a Legendre transform of $f(x)$, i.e., $\mathcal{V}(p) = f(x) - xp$. The overlap $p \equiv \mathbf{J} \cdot \mathbf{w}/N$ is related to the coupling field by $\frac{df(x)}{dx} = p$. Since the potential curve may lose its concavity, one has to solve numerically the saddle-point equations by fixing p and searching for compatible coupling field x (by using the secant method). If a solution of x is found for a given p , then we have $x = -\frac{d\mathcal{V}}{dp}$ at this value of p . Because $d = (1-p)/2$, x is also equal to $\frac{1}{2} \frac{d\mathcal{V}}{dd}$.

The derivative of the Franz-Parisi potential with respect to the overlap p is given by:

$$\frac{d\mathcal{V}(p)}{dp} = -\hat{p} + \alpha \frac{\partial}{\partial p} \int D\omega \int Dt H^{-1}(\tilde{t}) \int_{\tilde{t}}^{\infty} Dy \ln H(h(\omega, t, y)) \\ = -\hat{p} + \frac{\alpha}{\sqrt{(1-q)(1-r)}} \int D\omega \int Dt \mathcal{R}(\tilde{t}) \mathcal{R}(h(\omega, t, y = \tilde{t})). \quad (\text{B1})$$

Note that when $p \rightarrow 1$, r will get close to p but smaller than p , and $p' \simeq q$, which is observed in numerical simulations and can be understood from the definition of these order parameters. Therefore, in the limit $p = 1 - \epsilon \rightarrow 1$, the second term in the right-hand side of Eq. (B1) is $\alpha C_p \epsilon^{-1/2}$ with $C_p = \frac{1}{\sqrt{\pi(1-q)}} \int Dt \mathcal{R}(\tilde{t})$. The expression of \hat{p} as a function of ϵ can be deduced from Eq. (A9c). Using the fact that $p \rightarrow 1$ implies that $\hat{p} \rightarrow \infty$, and the identity

$\tanh(x) = 1 - 2e^{-2x}$ ($x \gg 0$), one finally gets $\hat{p} = \hat{r} - \frac{1}{2} \ln \frac{\epsilon}{2} + \frac{1}{2} \ln \int Dz_1 \int Dz_3 \frac{e^{\sqrt{\hat{q}-\hat{p}'}z_1 - \sqrt{\hat{p}'}z_3}}{\cosh(\sqrt{\hat{q}-\hat{p}'}z_1 + \sqrt{\hat{p}'}z_3)}$. In the above derivations, we have used the fact that $\frac{1-p}{1-r} = 1/2$ in the limit $p \rightarrow 1$ based on Eqs. (A9c) and (A9e). Taken together, one arrives at the slope of $\mathcal{V}(p)$ at $p = 1$:

$$\frac{d\mathcal{V}(p)}{dp}\bigg|_{p \rightarrow 1} = \frac{1}{2} \ln \frac{\epsilon}{2} + C + \alpha C_p \epsilon^{-1/2}. \quad (\text{B2})$$

Appendix C: Proof of $\mathcal{V}(p \rightarrow 1) = 0$

At $p = 1$, the Franz-Parisi potential can be expressed as:

$$\begin{aligned} \mathcal{V}(p) = & -p\hat{p} + p'\hat{p}' + \alpha \int D\omega \int Dt H^{-1}(\tilde{t}) \int_{\tilde{t}}^{\infty} Dy \ln H(h(\omega, t, y)) \\ & + \int D\mathbf{z} (2 \cosh \hat{a})^{-1} \left[e^{\hat{a}} \ln 2 \cosh(\hat{a}' + \hat{p} - \hat{p}') + e^{-\hat{a}} \ln 2 \cosh(\hat{a}' - \hat{p} + \hat{p}') \right]. \end{aligned} \quad (\text{C1})$$

Note that $h(\omega, t, y) = -\frac{1}{\sqrt{1-r}} (\sqrt{1-q}y + \sqrt{q}t) \rightarrow -\infty$ when $y > -\sqrt{\frac{q}{1-q}}t$. Hence the α -dependent term disappears. The last term becomes

$$\begin{aligned} & \int D\mathbf{z} (2 \cosh \hat{a})^{-1} \left[e^{\hat{a}} \ln 2 \cosh(\hat{a}' + \hat{p} - \hat{p}') + e^{-\hat{a}} \ln 2 \cosh(\hat{a}' - \hat{p} + \hat{p}') \right] \\ & = \int Dz_1 \int Dz_3 (2 \cosh \hat{a})^{-1} \left[e^{\hat{a}} (\sqrt{\hat{p}'}z_3 + \hat{p} - \hat{p}') + e^{-\hat{a}} (-\sqrt{\hat{p}'}z_3 + \hat{p} - \hat{p}') \right] \\ & = \hat{p} - \hat{p}' + \hat{p}' \left[1 - \int Dz_1 \int Dz_3 \tanh^2(\sqrt{\hat{q}-\hat{p}'}z_1 + \sqrt{\hat{p}'}z_3) \right] \\ & = \hat{p} - q\hat{p}'. \end{aligned} \quad (\text{C2})$$

Collecting the above results, one arrives at $\mathcal{V}(p \rightarrow 1) = 0$.

-
- [1] M. Mézard and A. Montanari, *Information, Physics, and Computation* (Oxford University Press, Oxford, 2009).
 - [2] F. Krzakala, A. Montanari, F. Ricci-Tersenghi, G. Semerjian, and L. Zdeborova, *Proc. Natl. Acad. Sci. USA* **104**, 10318 (2007).
 - [3] W. Krauth and M. Mézard, *J. Phys. (France)* **50**, 3057 (1989).
 - [4] A. Engel and C. V. den Broeck, *Statistical Mechanics of Learning* (Cambridge University Press, Cambridge, England, 2001).
 - [5] T. Hosaka, Y. Kabashima, and H. Nishimori, *Phys. Rev. E* **66**, 066126 (2002).
 - [6] A. Lage-Castellanos, A. Pagnani, M. Weigt, and R. Zecchina, *J. Stat. Mech.: Theory Exp* p. P10009 (2009).
 - [7] A. L. Blum and R. L. Rivest, *Neural Networks* **5**, 117 (1992).
 - [8] H. M. Köhler, *J. Phys. A: Math. Gen.* **23**, L1265 (1990).
 - [9] H. K. Patel, *Z. Phys. B* **91**, 257 (1993).
 - [10] M. Bouten, L. Reimers, and B. V. Rompaey, *Phys. Rev. E* **58**, 2378 (1998).
 - [11] A. Braunstein and R. Zecchina, *Phys. Rev. Lett* **96**, 030201 (2006).
 - [12] T. Shinzato and Y. Kabashima, *J. Phys. A: Math. Theor.* **41**, 324013 (2008).
 - [13] H. Huang and H. Zhou, *J. Stat. Mech.: Theory Exp* p. P08014 (2010).
 - [14] H. Huang and H. Zhou, *Europhys. Lett* **96**, 58003 (2011).
 - [15] R. C. Alamino, J. P. Neirotti, and D. Saad, *Phys. Rev. E* **88**, 013313 (2013).
 - [16] H. Horner, *Z. Phys. B* **86**, 291 (1992).
 - [17] T. Obuchi and Y. Kabashima, *J. Stat. Mech.: Theory Exp* p. P12014 (2009).
 - [18] H. Huang, K. Y. M. Wong, and Y. Kabashima, *J. Phys. A: Math. Theor.* **46**, 375002 (2013).
 - [19] E. Gardner and B. Derrida, *J. Phys. A: Math. Gen.* **21**, 271 (1988).
 - [20] S. Franz and G. Parisi, *J. Phys. I France* **5**, 1401 (1995).
 - [21] S. Franz and G. Parisi, *Phys. Rev. Lett* **79**, 2486 (1997).

- [22] S. Franz and G. Parisi, *Physica A* **261**, 317 (1998).
- [23] J. Nocedal and S. Wright, *Numerical Optimization* (Springer, Berlin, 2006).
- [24] C. Di, A. Montanari, and R. Urbanke, in *Proc. IEEE Int. Symp. Information Theory* (Chicago, 2004), p. 102.
- [25] L. Zdeborová and M. Mézard, *Phys. Rev. Lett* **101**, 078702 (2008).
- [26] L. Zdeborová and F. Krzakala, *Phys. Rev. B* **81**, 224205 (2010).
- [27] T. Uezu and K. Nokura, *Prog. Theor. Phys.* **95**, 273 (1996).
- [28] S. Franz and G. Parisi, *J. Stat. Mech.: Theory Exp* p. P11012 (2013).
- [29] D. Achlioptas, A. Naor, and Y. Peres, *Nature* **435**, 759 (2005).
- [30] R. Gütig and H. Sompolinsky, *Nat. Neurosci* **9**, 420 (2006).
- [31] R. Rubin, R. Monasson, and H. Sompolinsky, *Phys. Rev. Lett* **105**, 218102 (2010).
- [32] C. Baldassi, A. Braunstein, and R. Zecchina, *J. Stat. Mech.: Theory Exp* p. P12013 (2013).